Linear Regression as a 1-Variable Optimization Exercise

Ken Constantine
Current affiliation: Eastern Nazarene College
Fall affiliation: Taylor University

Abstract:
Derivation of the least squares line for a set of bivariate data entails minimizing a function of two variables, say the line's slope and intercept. Imposing the requirement that the line pass through the mean point for the data reduces this problem to a 1-variable problem easily solved as a single-variable Calculus exercise. The solution to this problem is, in fact, the solution to the more general problem. We illustrate with a dataset involving charitable donations.

EXAMPLE:
The population for our example is from the *Chronicle of Philanthropy*, 1 May 2003, page 12. There are 100 U. S. metropolitan areas and for each is given: the number of itemized tax returns filed, the average discretionary income for those returns, and the average charitable donation amount for those returns. See the URL

http://alpha2.enc.edu/~constant/library/statistics/data/charity/charity.xls

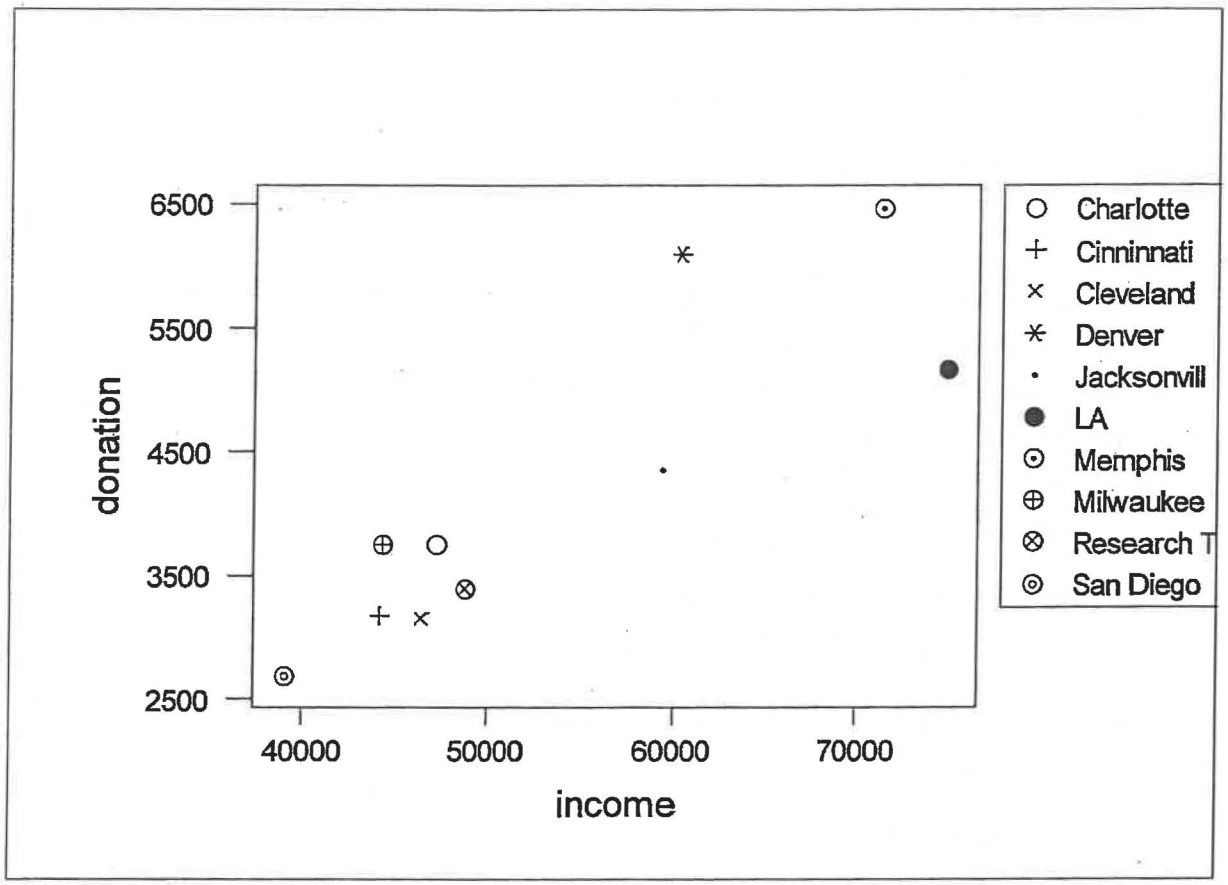
for an Excel spreadsheet providing this information for the full population.

From this population a sample of 10 cities was selected by simple random sampling. Those cities and their data are as follows:

city	**income**	**donation**
Charlotte	47,262	3,747
Cinninnati	44,229	3,163
Cleveland	46,425	3,141
Denver	60,326	6,094
Jacksonville	59,444	4,356
LA	74,960	5,169
Memphis	71,335	6,464
Milwaukee	44,396	3,749
Research Triangle	48,783	3,383
San Diego	39,086	2,680

Descriptive Information for these Data:
We begin with descriptive statistics for these data.

(graph)



(numerical measures)

Variable	N	Mean	Median	StDev	Min	Max	Q1	Q3
Income	10	53,625	48,022	12,239	39,086	74,960	44,354	63,078
Donation	10	4,195	3,748	1,302	2,680	6,464	3,157	5,400

Correlation of income and donation = 0.870

least squares

The classic criterion for a line which "best" fits a dataset such as ours is that the sum of the squared vertical distances $y_i - [a + bx_i]$ between line and data points be minimized. If the line is given by y=a+bx, the sum to be minimized is

$$D(a,b) = \sum_{i=1}^{n} (y_i - [a + bx_i])^2 .$$

The minimization of D with respect to the variables a and b is customarily a multivariable Calculus or Linear Algebra problem.

An alternative approach is to impose the ad hoc but intuitively reasonable assumption that the line must pass through the mean point $(\bar{x}, \bar{y})$. (It turns out that the full solution passes through this point and so our ad hoc assumption will in fact generate the full-fledged least squares line.) Under our constraint, the equation of our line is $y = \bar{y} + b(x - \bar{x})$ and so the sum to be minimized is a function of the slope b alone:

$$d(b) = \sum_{i=1}^{n} (y_i - [\bar{y} + b(x_i - \bar{x})])^2 .$$

It is worth noting that d is a quadratic and convex function of b:

$$d(b) = \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2b \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

so that the minimum is easily found via one-variable Calculus (or even pre-Calculus) methods. The minimizer is

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} .$$

pedagogy

As a classroom exercise, there are different approaches available which might be well-suited to different students and objectives. If mathematical maturity were the goal, a general derivation (as above) might be suitable. If notation were to be downplayed, specific data could be used if that were better-suited to a group of students. For the data in our example, the objective function is

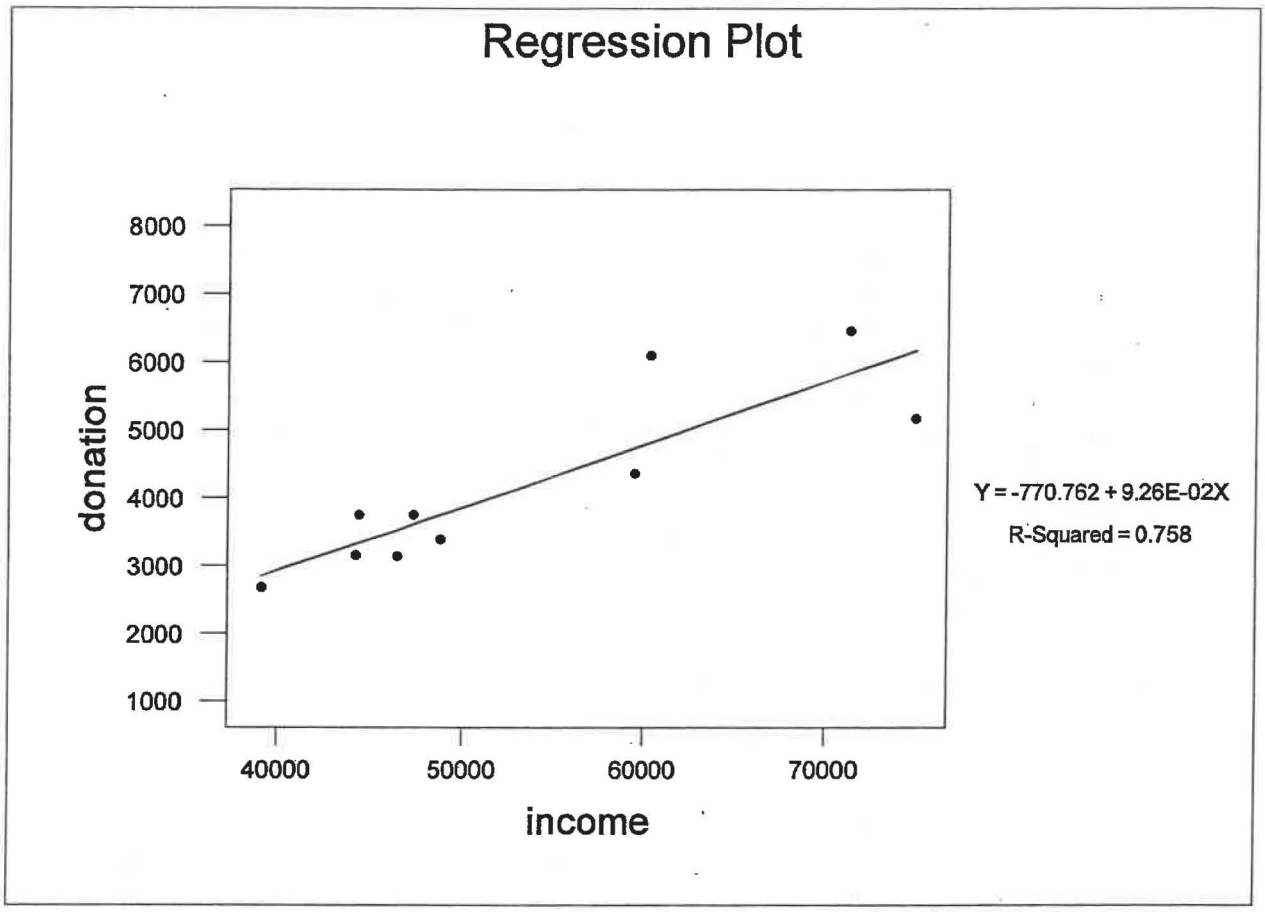
$$d(b) = 15{,}259{,}286 - 249{,}722{,}495\, b + 1{,}348{,}539{,}230\, b^2$$

and the minimizer is $\hat{b} = .0926$.

A DERIVE worksheet for our example may be downloaded from the URL
http://alpha2.enc.edu/~constant/library/statistics/data/charity/charity.dfw

EXAMPLE

The following output from Minitab shows that the graph of the above line.



additional issues:

There are other topics which this exercise can serve to introduce. One such issue is the meaning of the coefficient of determination R^2. Another topic is the interpretation of the least squares line and the distinction between relationship and causation. Finally, there are a number of application issues available for discussion such as the reasons for differences between cities' donations and income.

[The following is an insert in an abstract algebra text, appearing after a section of some results from the elementary theory of equations and immediately before the presentation of Kronecker's theorem.]

A Pseudo-History of Number Systems

(or "The Way Mathematicians Wish the Number Systems Had Been Invented Because It Makes a More Cohesive Story")

Once upon a time, sometime after the beginning, there were the natural numbers -- because people wanted to count things and other people. Then a trouble-maker got up in a meeting of the Neolithic Mathematical Society and said, "I can form a lot of equations using integer coefficients, such as $2x = 3$, that I cannot solve." So the folks at the meeting made a rational decision to form a committee, and the committee reported back that there should be fractions such as *3/2* and *1/4* and even $243 / 19762$ so that equations such as $ax = b$ could be solved. But at the end of the report was the comment: "On a negative note, this still doesn't enable us to solve equations such as $x + 5 = 2$." So the meeting decided to form another committee (including, of course, several members of the first committee), which went out and studied the problem for a long time and finally reported back that there should be "negative" numbers like -34 and -3 and even $-243 / 19762$ so that all linear equations with natural number coefficients could be solved. And the mathematicians all rejoiced, but their students groaned because this meant that there would be a lot more homework for them to do.

So the students got together and brought the equation $x^2 = 2$ to class one day and asked their teacher to solve it in this wonderful system which the second committee had given them to do their homework, and the teacher was stumped. But he went home and thought about the problem all night and came to class the next day and said, "We need more numbers, a whole line full, that I have decided to call the real numbers because I worked real hard to figure them out. They include $\sqrt{2}$, which is a solution to your equation, and lots of other numbers -- more than you need to solve all of the equations $x^2 = q$, where q is a positive fraction. (There are some that are beyond what my dentist can use which you can name to suit yourselves.)" But the head of his department heard of the teacher's work and thought the teacher had become irrational, so the teacher lost his job and was replaced by an even smarter teacher. When she heard why her predecessor was fired, she asked her students, "What if q is negative? Can you solve the equation $x^2 = -1$?" This question greatly troubled her students and was too complex for most of them.

However, one student, who was studying to be an engineer, said, "If my last teacher can invent a number whose square is 2, why don't we invent another number whose square is -1? Even if there isn't any such thing, we can imagine it." So they called it an imaginary number, and they solved twice as many equations with it; however, the engineer was shunned by his classmates because now there were more homework problems than ever before. So he switched his major to Accounting where he didn't have to deal with complicated numbers anymore. But everybody still had lots of homework.

Moral #1: If you want to solve an equation and can't find a number that solves it, invent one.

Moral #2: If you want to invent a whole new number system, be sure to give the new numbers names like "negative" or "irrational" or "imaginary" so no one will think you're serious. If you wait long enough and your system catches on, you will be considered a brilliant innovator.

Moral #3: If you don't feel very innovative today, you can ignore Morals #1 and #2 and move to Section 6.5 where the new systems you need to solve equations in $F[x]$ (where F is a field) will be developed.

(-- as told to Richard Laatsch)